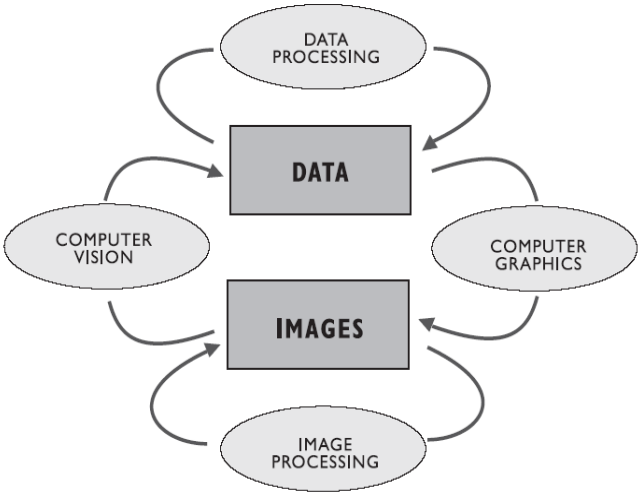
15. Einführung in Computer Vision

Was ist Visual Computing?

In **Visual Computing** unterscheiden wir generell 3 Teilbereiche: Computergrafik, Computer Vision (Maschinelles Sehen, Bildverstehen, Mustererkennung) und Bildverarbeitung. Alle drei Teilbereiche arbeiten zusammen, wobei wir zwischen Daten und Bildern unterscheiden. Daten sind allgemein alle Arten von digitalen Informationen, die in der Datenverarbeitung bearbeitet werden.



Wenn wir aus den numerischen Daten bildhafte Daten erzeugen wollen, benötigen wir dazu **Computergrafik**, die sich mit der computergestützten Erzeugung von Bildern befasst. Dazu gehören zum Beispiel die Generierung von Balken- und Tortendiagrammen aus Daten (Visualisierung) aber auch die Erzeugung von Grafiken (computergenerierte Bilder), deren Bestandteile sich zweidimensional in der Ebene oder im 3D Raum beschreiben lassen. Ein Teilbereich der Computergrafik versucht aus Modellen möglichst realistische Bilder zu generieren, die zum Beispiel durch Aneinanderreihung von Teilbildern Animationsfilme erzeugen.

Computer Vision versucht das Gegenteil: aus realen Bildern die Semantik zu extrahieren, um damit zum Beispiel Modelle generieren zu können. Anders gesagt beschäftigt sich Computer Vision mit dem Problem, Sehvorgänge in der realen, dreidimensionalen Welt nachzubilden. Dazu gehört die räumliche Erfassung von Gegenständen und Szenen, das Erkennen von Objekten, die Interpretation von Bewegungen, autonome Navigation, das mechanische Aufgreifen von Dingen (durch Roboter) usw. Computer Vision entwickelte sich ursprünglich als Teilgebiet der „Künstlichen Intelligenz“ (Artificial Intelligence, kurz „AI“) und die Entwicklung zahlreicher AI-Methoden wurde von visuellen Problemstellungen motiviert.

Ein Teilgebiet der Computer Vision ist die **Mustererkennung**, die sich allgemein mit dem Auffinden von „Mustern“ in Daten und Signalen beschäftigt. Typische Beispiele aus diesem Bereich sind etwa die Unterscheidung von Texturen oder die optische Zeichenerkennung (Optical Character Recognition, kurz „OCR“). Diese Methoden betreffen aber nicht nur Bilddaten, sondern auch Sprach- und Audiosignale, Texte, Börsenkurse, Verkehrsdaten, die Inhalte großer Datenbanken u.v.m. Statistische und syntaktische Methoden spielen in der Mustererkennung eine zentrale Rolle.

Bildverarbeitung (engl. Image Processing) hingegen bleibt auf der Bildebene, sie versucht Bilder so zu behandeln, dass sie für bestimmte Aufgabenstellungen besser geeignet sind. Es werden also Bilder, beispielsweise Fotografien oder Einzelbilder aus Videos, verarbeitet, um schärfere oder komprimierte Bilder zu erzeugen. Das Ergebnis der Bildverarbeitung ist wiederum ein Bild. In den meisten Fällen werden Bilder als zweidimensionales Signal betrachtet, sodass Methoden aus der Signalverarbeitung angewandt werden können. Bildverarbeitung ist zu unterscheiden von der Bildbearbeitung, die sich mit der Manipulation von Bildern zur anschließenden Darstellung beschäftigt, zum Beispiel mithilfe von Adobe Photoshop. Bildbearbeitung passiert daher interaktiv und mit visueller Kontrolle, Bildverarbeitung folgt mathematischen, algorithmischen Prozessen. Beispielgebiete für Bildverarbeitung sind Entfernung von Bildrauschen, Verbesserung der Bildqualität, künstlerische Effekte, etc.

Definition Computer Vision

"Computer Vision describes the automatic deduction of the structure and the properties of a (possible dynamic) three-dimensional world from either a single or multiple two-dimensional images of the world. The images may be monochromatic (i.e., "black and white") or colored, they may be captured by a single or multiple cameras, and each camera may be either stationary or mobile " - Vishvjit S. Nalwa: A guided tour of computer vision. Addison-Wesley 1993.
Eine industrienähere Version von Microsoft Research: *“Computer Vision is an exciting new research area that studies how to make computers efficiently perceive, process, and understand visual data such as images and videos. The*

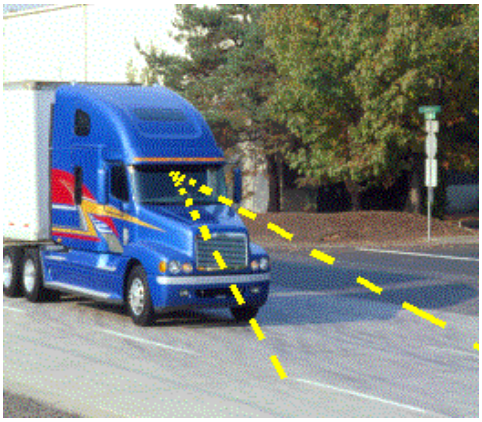
ultimate goal is for computers to emulate the striking perceptual capability of human eyes and brains, or even to surpass and assist the human in certain ways". – Microsoft Research.

Computer Vision ist allgemein ein Bereich, der Methoden zur Aufnahme, Verarbeitung, Analyse und zum Verstehen von Bildern beinhaltet und sich mit bildhaften Daten der realen Welt befasst, um numerische oder symbolische Informationen zu produzieren, z.B. in Form von Entscheidungen. Eines der Themen dieses Forschungsgebiets ist es, die Fähigkeit des menschlichen Sehens zu reproduzieren, sodass Bilder auf elektronischem Wege wahrgenommen und verstanden werden können. Das Verstehen von Bildern kann auch als das Entwirren von symbolischer Informationen in Bilddaten angesehen werden. Dies erfolgt durch die Verwendung von Modellen, die mit Hilfe der Geometrie, Physik, Statistik und Lerntheorie konstruiert werden.

Der Anwendungsbereich der Computer Vision reicht von industriellen Bildverarbeitungssystemen (z.B. das Erkennen von vorbeifahrenden Flaschen auf einem Förderband) bis hin zu Forschungsgebieten im Bereich Artificial Intelligence (z.B. Computer oder Roboter, die ihre Umwelt verstehen können). Auch im Bild links ist ein typischer Anwendungsbereich der Computer Vision zu sehen: die automatische Detektion von Gesichtern zur verbesserten Bildaufnahme von Digitalkameras. Die Bereiche Computer Vision und Machine Vision überdecken sich wesentlich. Computer Vision befasst sich mit der Kerntechnologie der automatisierten Bildanalyse, welche in vielen Bereichen verwendet wird. Machine Vision bezieht sich üblicherweise auf Prozesse, welche die automatisierte Bildanalyse mit anderen Methoden und Technologien kombinieren, z.B. automatische Überprüfung und Steuerung eines Roboters im industriellen Anwendungsbereich.

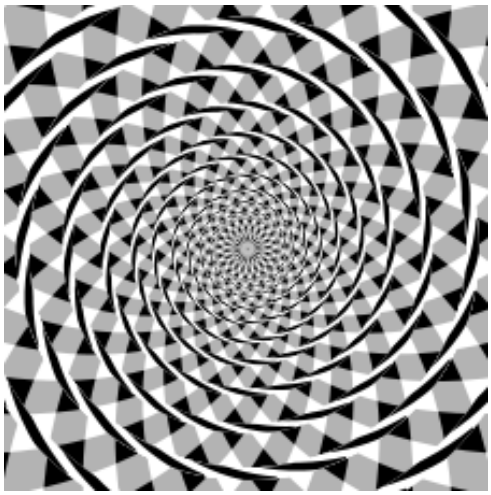
In der technologischen Disziplin werden Theorien und Modelle von Computer Vision für die Konstruktion von Computer Vision Systemen angewendet. Anwendungsbeispiele für solche Systeme sind:

- Kontrollprozesse (z.B. Industrieroboter)
- Navigation (z.B. autonome Fahrzeuge oder mobile Roboter)
- Erkennen von Ereignissen (z.B. Zählen von Menschen oder visuelle Überwachung)
- Informationsorganisation (z.B. Katalogisierung von Datenbanken oder Bildern und Bildsequenzen).
- Modellierung von Objekten und Umgebungen (z.B. medizinische Bildanalyse oder topographische Modellbildung)
- Interaktion (z.B. Eingabe für ein Gerät für die Interaktion eines Menschen mit einem Computer)
- Automatische Überprüfung (z.B. Anwendungen im Bereich der Produktion)



Teilgebiete der Computer Vision beinhalten die Rekonstruktion von Szenen, die Detektion von Ereignissen, Videotracking, Objekterkennung, Maschinelles Lernen, Katalogisierung, Bewegungsabschätzung und Bildrestauration.

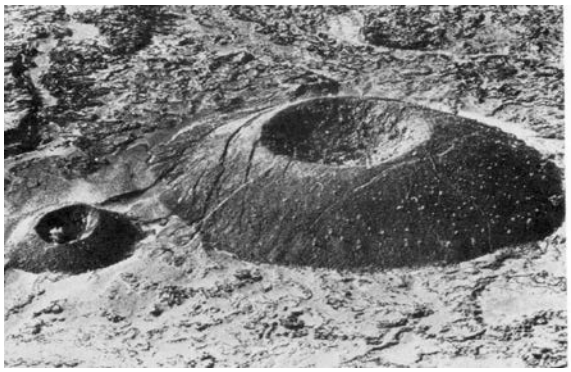
Mensch vs. Computer Vision



In der Computer Vision wurde immer untersucht, wie die Fähigkeit des menschlichen Sehens reproduziert werden kann. Es muss jedoch bedacht werden, dass diese Fähigkeit nicht nur das Ergebnis Millionen Jahre langer Evolution ist, sondern auch, dass das menschliche Sehen Fehlbarkeiten aufweist. Dies kann durch die altbekannte optische Täuschung der Fraser-Spirale aufgezeigt werden (siehe Abbildung links), die in Wirklichkeit keine ist.

In der Computer Vision versuchen wir die Struktur und die Eigenschaften der dreidimensionalen Welt herzuleiten. Dies erfolgt nicht nur durch die Verwendung von geometrischen Eigenschaften, sondern auch unter Einbeziehung der Materialeigenschaften und der Beleuchtung. Beispiele für geometrische Eigenschaften sind Formen, Größen und Lage der Objekte. Materialeigenschaften sind z.B. die Helligkeit oder Dunkelheit von Oberflächen, deren Farbe, Texturen und Materialzusammensetzung.

Es stellt sich die Frage, wieso Forscher der Computer Vision nicht einfach Systeme konstruieren, die das visuelle System des Menschen emulieren, besonders unter Einbeziehung der Literatur der Neurophysiologie, Psychologie und Psychophysik. Ein guter Grund dafür, weshalb Computer Vision Forscher dies unterlassen, ist das lediglich spekulative und geringe Wissen über die Vorgänge des visuellen Systems, die nach der Reizaufnahme durch das menschliche Auge passieren. Für die gängigsten Tätigkeiten ist das menschliche Sehen ausreichend, diese Zweckdienlichkeit sollte aber nicht zu dem Rückschluss der Unfehlbarkeit führen. Die Fehlbarkeit des menschlichen Sehens wird durch die Existenz der optischen Täuschungen, Mehrdeutigkeiten sowie Inkonsistenzen aufgezeigt (siehe Abbildung oben).

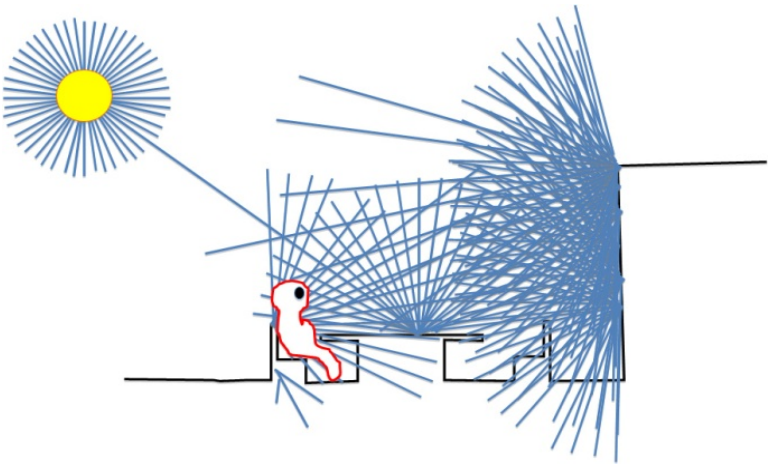


Die zahlreichen Beispiele für optische Täuschungen, Mehrdeutigkeiten und Inkonsistenzen führen zu der Annahme, dass die Manipulationen unseres visuellen Systems auf dessen Mangel an „Realismus“ beruhen. Wenn wir nicht an das glauben können, was wir sehen, was können wir glauben? Dieses absolute Vertrauen in unsere perzeptuellen Fähigkeiten ist nicht gänzlich gerechtfertigt. Die Fotografie auf der linken Seite stellt eine natürliche Szene dar, welche Lavakegel mit Kratern an deren Spitzen zeigt. Es macht den Eindruck, dass es keinen Grund für eine Täuschung gibt, da eine reale Szene vorliegt. Wird die Fotografie jedoch um 180° gedreht, werden die Kegel zu Kratern und die Krater werden zu Kegel.

Die verschiedenen optischen Täuschungen, Mehrdeutigkeiten und Inkonsistenzen sind mehr als nur eine kuriose Spielerei. Sie geben uns wertvolle Einblicke in die Natur des menschlichen Sehens und lassen zusätzlich folgende wichtige Frage aufkommen: Ist das menschliche Sehen nur eine kontrollierte Halluzination? Können wir durch netzhautgenerierte Bilder mehr Rückschlüsse ziehen als durch künstlich mit Hilfe der Geometrie und Physik erstellte Bilder? Helmholtz vertritt die Ansicht: “*… dass wir stets solche Objecte als im Gesichtsfelde vorhanden uns vorstellen, wie sie vorhanden sein müssten, um unter den gewöhnlichen normalen Bedingungen des Gebrauchs unserer Augen denselben Eindruck auf den Nervenapparat hervorzubringen*” (*Handbuch der physiologischen Optik*, Helmholtz 1910, Seite 428). Das Ergebnis ist, dass Menschen etwas „sehen“ können, das nicht existiert (d.h. halluzinieren) oder etwas nicht wahrnehmen, das sie umgibt (d.h. übersehen). Während wir die (in manchen Bereichen) mangelhafte Leistung des visuellen Systems der Menschen akzeptieren, sind wir im Bezug auf die Leistung von Maschinen nicht so nachsichtig. Daher müssen wir uns selber die Frage stellen: möchten wir wirklich, dass Maschinen so sehen können wie wir?

Plenoptische Funktion

Zuerst stellen wir uns die Frage, was potentiell gesehen werden kann. Welche Informationen über die Welt kann ein raumfüllendes Licht beinhalten? Räume werden mit einem dichten Spektrum an Lichtstrahlen unterschiedlicher Intensität gefüllt. Die Menge an Strahlen, die durch jeden Punkt im Raum gehen, wird in der Mathematik als „Strahlenbündel“ bezeichnet. Dieses Bündel wird von Leonardo da Vinci „strahlende Pyramiden“ bezeichnet:



“The air is full of an infinite number of radiant pyramids caused by the objects located in it. These pyramids intersect and interweave without interfering with each other during the independent passage throughout the air in which they are infused.” (*The Notebooks of Leonardo da Vinci*, Leonardo da Vinci und Irma A. Richter, Oxford University Press, 1980).

Wir wollen nun die Parameter zur Beschreibung beleuchteter Umgebungen betrachten. Im ersten Schritt nehmen wir eine Schwarzweißfotografie einer Lochkamera an, die uns darüber Auskunft gibt, wie groß die Intensität des Lichtes von einem einzigen Blickpunkt aus zu einem bestimmten Zeitpunkt, gemittelt über die Wellenlängen des Spektrums des sichtbaren Lichtes, ist. Diese Lichtintensitätsverteilung P kann beim Durchgang der Strahlenbündel durch die Linse gemessen und anhand von Kugelkoordinaten $P(\theta,\varphi)$ oder kartesischen Koordinaten $P(x,y)$ parametrisiert werden.

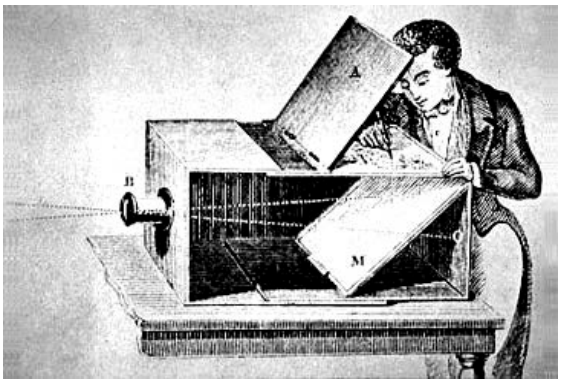
Ein Farbbild enthält zusätzliche Informationen, welche die Veränderung der Lichtintensität bezüglich der Wellenlänge λ berücksichtigt (daher $P(\theta, \varphi, \lambda)$). Ein Farbvideo oder Film erweitert die Informationen um die Dimension der Zeit t: $P(\theta, \varphi, \lambda, t)$. Schlussendlich zeigen farbige holografische Filme die komplette wahrnehmbare Lichtintensität von jeder

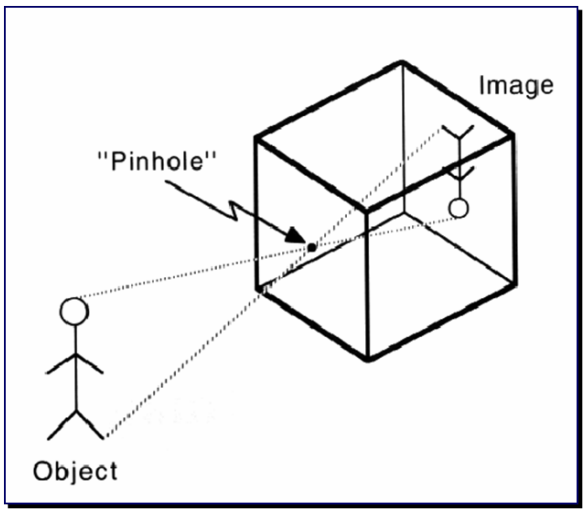
Betrachtungsposition V_x, V_y, and V_z an: $P(\theta, \varphi, \lambda, t, V_x, V_y, V_z)$. Solch eine vollständige Repräsentation impliziert die Beschreibung aller möglichen Bilder, die von einem bestimmten Raum-Zeit Stück der Welt aufgenommen werden können (unter Vernachlässigung der Polarisation und momentanen Phase des einfallenden Lichtes). Es ist zu beachten, dass die plenoptische Funktion keine zusätzlichen Parameter zur Spezifizierung der "Blickrichtung" des Auges benötigt. Die Veränderung der Blickrichtung ohne Veränderung der Position des Auges hat keine Auswirkung auf die Verteilung des Lichtes eines Strahlenbündels beim Auftreffen auf die Pupille. Nur die relative Einfallposition des Lichtes auf der Netzhaut wird dadurch verändert.

Die Messung der plenoptischen Funktion erfolgt durch die imaginäre Platzierung eines idealen Auges an jeder möglichen (V_x,V_y,V_z) Position und beinhaltet die Messung der Intensität der Lichtstrahlen, für jeden möglichen Einfallswinkel (θ,φ), für jede Wellenlänge λ, zu jedem Zeitpunkt t, die durch das Zentrum der Pupille gehen. Die Winkel (θ,φ) können immer relativ zu einer optischen Achse, die parallel zur V_z Achse liegt, beschrieben werden. Die resultierende Funktion hat die Form: $p = P(\theta, \varphi, \lambda, t, V_x, V_y, V_z)$.

Camera Obscura / Lochkamera

Das einfachste Prinzip einer Kamera ist die so genannte Lochkamera, die bereits im 13. Jahrhundert als „Camera obscura“ (Latein; „camera": Raum + „obscura": dunkel = abgedunkelter Raum) bekannt war. Sie fanden Anwendung in der Malerei um geometrisch richtige, detailgetreue Abbildungen malen zu können, sowie in der Unterhaltungsbranche, um unerkannt die Außenwelt beobachten zu können (erste „Spycam“). Die Lochkamera war jene Erfindung, die zur Fotografie führte. Sie hat zwar heute keinerlei praktische Bedeutung mehr (eventuell als Spielzeug), aber sie dient als brauchbares Modell, um die wesentlichen Elemente der optischen Abbildung zu beschreiben.





Die Lochkamera besteht aus einer geschlossenen Box mit einer winzigen Öffnung an der Vorderseite („Pinhole“) und der Bildebene an der gegenüberliegenden Rückseite (siehe Abbildung links). Lichtstrahlen, die von einem Objektpunkt vor der Kamera ausgehend durch die Öffnung einfallen, werden geradlinig auf die Bildebene projiziert, wodurch ein verkleinertes und seitenverkehrtes Abbild der sichtbaren Szene entsteht. Durch den Einsatz von Spiegeln kann das projizierte Bild in die richtige Position gedreht werden. Eine portable Version der Lochkamera stellte eine Kiste mit einem angewinkelten Spiegel dar, mit denen ein Bild in richtiger Ausrichtung auf ein durchscheinendes Papier, welches auf einer Glasoberfläche liegt, projiziert wird. Die Lochkamera kann als einfache Kamera ohne Linse und einer einzigen, kleinen Apertur (Öffnung) angesehen werden.